o Review and identify variables to be used for segmentation

o Review clustering techniques and select two

o Perform the segmentation

o Statistically evaluate the results and compare the two techniques

o Group all the customers into the decided segments

o Provide a detailed overview (i.e., the number of segments, cross-segment differences, segment-specific profiles, etc.) of the selected segmentation results, expressly differentiating between ‘high’ and ‘low’ value segments; make sure to provide sound justification for why you consider some segments to be ‘high value’ while others to be ‘low value’

Abstract

Introduction (Maybe)

**Attribute Identification**

Before the attribute identification was hypothesized and decided upon using a cleaned-up form of the *Merrimack\_Customers.xls* dataset provided, a few final modifications to the set were executed. There were a couple of missed opportunities because of due diligence and those were addressed. *CarValue* contained a -1 in the place of any customer who did not own or lease a vehicle. This field was converted to *NoCar* to maintain the categorical effectiveness and intuitiveness of the data. The created variable *AvgPhoneBill* had missing values where the computation could not be executed (When tenure was 0 the division could not be done). These situations were converted to be equal to *VoiceOverTenure* as that is also the average in these situations. Lastly *CarValue* has a purposely empty value where -$1000 was converted to NaN. The NaN fields were converted to 0 to be more sensible. Someone who does not own a car and someone owning a car worth $0 reflect the real world (a broken old rust bucket is at least worth a few cents!).

To attempt a segmentation of customers based on probability of retention, three variables were decided upon. Two were existing in the current set and one was a combination of other variables to make the segmentation more pointed. The question was poised: ***What variables in the data when combined may cause clear clustering, separating customers who stay with the company from those who have a higher possibility of leaving?***

Appendix example one is a list of natural variable groupings created when executing data exploration due diligence. From the career demographic, employment length was chosen as a possible indicator. The mean, median and kurtosis of this variable in the data set showed malformed normality. To normalize the set a variable was adapted from *EmploymentLength* and given the label *LogEmployment*. It is the log of the *EmploymentLength + 1* as the logarithm of 0 (people without employment) is not a real number. The data set ranged from 0-52 so it was altered to be 1-53. To return from the calculation one only must take the inverse log of the value and subtract 1. (It is worth nothing that *EmploymentLength* shared very small ANOVA p-values with many other attributes in the data set. See example 3.)

The second variable chosen was created from *TotalOverTenure*. The attribute was created during due diligence to reflect all the money spent by a given customer record. The attribute used for cluster analysis was labeled *LogAvgTotal*. This attribute was created by dividing the total over tenure by the tenure, adding 1 and taking the logarithm. This procedure normalized the data completely and created an attribute that shares correlations or produces small-valued in ANOVA tests with many lifestyle and financial demographic variables (example 4). Neither of the two selected variables shares an association with each other or with the last variable, *PhoneCoTenure.*

*PhoneCoTenure* is a numeric variable that ranges from 0 to 72 and indicates the months a customer has belonged to the company. The purpose of this project is to find customers that will hang around. There is no better customer demographic attribute to describe that than *PhoneCoTenure.* This attribute adhered to a normal distribution without modification.

Clustering Technique Selection

The first clustering technique utilized was k-means clustering

Some Other (Hierarchal?)

Segmentation and Results

Customer Groups (detailed overview)

Conclusion

Attribute Identification

The goal of the data exploration clustering analysis is to produce a customer profile that segments customers based on chance of retention.

Appendix

Example 1:

***Natural Demographic Grouping***

* **Geographic**
  + Region
  + TownSize
* **Age**
  + Age
* **Gender**
  + Gender
* **Career**
  + JobCategory
  + UnionMember
  + EmploymentLength
  + Retired
  + Education
* **Lifestyle**
  + MaritalStatus
  + HouseholdSize
  + CommuteTime
  + PoliticalPartyMem
  + Votes
  + ActiveLifestyle
  + EBilling
  + TVWatchingHours
  + OwnsPC
  + OwnsGameSystem
  + OwnsMobileDevice
  + OwnsFax
  + NewsSubscriber
  + HasPets
* **Financial**
  + LoanDefault
  + HomeOwner
  + CarsOwned
  + CarOwnership
  + CarValue
  + CreditCard
  + CardTenure
  + CardItemsMonthly
  + CardSpendMonth
  + InHeavyDebt
* **Current** **Customer** **Profile**
  + PhoneCoTenure
  + VoiceLastMonth
  + VoiceOverTenure
  + EquipmentRental
  + EquipmentLastMonth
  + EquipmentOverTenure
  + CallingCard
  + WirelessData
  + DataLastMonth
  + DataOverTenure
  + TotalOverTenure
  + AvgPhoneBill
  + Multiline
  + VM
  + Pager
  + Internet
  + CallerID, CallWait, CallForward, ThreeWayCalling

Example 2:

LogEmployment, float64

count 5000.000000

mean 1.895185

std 1.071595

min 0.000000

25% 1.098612

50% 2.079442

75% 2.772589

max 3.970292

Normality Snapshot:

Mean: 1.89518478841

Median: 2.0794415416798357

Skew: -0.347851464019

Kurtosis: -0.874810770554

Example 3:

Possible Associations (EmploymentLength Comparison):

TownSize: P-value (ANOVA) ->5.6006469535e-05

JobCategory: P-value (ANOVA) ->1.44267168586e-321

UnionMember: P-value (ANOVA) ->0.0137349412352

Retired: P-value (ANOVA) ->9.03769582655e-319

LoanDefault: P-value (ANOVA) ->5.39457765294e-125

ActiveLifestyle: P-value (ANOVA) ->5.76000479606e-33

EquipmentRental: P-value (ANOVA) ->3.1395088607e-35

EquipmentLastMonth: P-value (ANOVA) ->3.1395088607e-35

CallingCard: P-value (ANOVA) ->5.97393043958e-98

WirelessData: P-value (ANOVA) ->9.9462551089e-09

Multiline: P-value (ANOVA) ->1.1087107191e-56

VM: P-value (ANOVA) ->1.12822622171e-08

Pager: P-value (ANOVA) ->7.14330696407e-10

Internet: P-value (ANOVA) ->4.38111765529e-28

CallWait: P-value (ANOVA) ->0.00933686815736

CallForward: P-value (ANOVA) ->0.038399708387

ThreeWayCalling: P-value (ANOVA) ->0.0437036243162

EBilling: P-value (ANOVA) ->3.63472966578e-27

OwnsPC: P-value (ANOVA) ->1.65039111883e-30

OwnsMobileDevice: P-value (ANOVA) ->1.99371337421e-45

OwnsGameSystem: P-value (ANOVA) ->4.67070666913e-55

OwnsFax: P-value (ANOVA) ->0.00448744464508

NewsSubscriber: P-value (ANOVA) ->2.15052442545e-159

HouseholdIncome: P-value (ANOVA) ->2.90457792621e-159

Education: P-value (ANOVA) ->1.59368123993e-50

InHeavyDebt: P-value (ANOVA) ->0.0430093592582

EquipmentOverTeunre: P-value (ANOVA) ->3.1395088607e-35

Example 4:

58: LogAvgTotal, float64

count 5000.000000

mean 3.402475

std 0.802680

min 0.000000

25% 2.832901

50% 3.491952

75% 4.001514

max 6.026790

Normality Snapshot:

Mean: 3.40247451289

Median: 3.491951744930619

Skew: -0.337234277062

Kurtosis: -0.242204525748

Possible Associations:

TotalAvgBill: R-Squared (Correlation) ->0.883561358379

JobCategory: P-value (ANOVA) ->3.03135112279e-19

Retired: P-value (ANOVA) ->3.75027948101e-41

LoanDefault: P-value (ANOVA) ->8.42952603844e-21

MaritalStatus: P-value (ANOVA) ->3.42526928988e-18

HomeOwner: P-value (ANOVA) ->8.86334046129e-11

Votes: P-value (ANOVA) ->0.0362212206346

CreditCard: P-value (ANOVA) ->1.02263631799e-16

ActiveLifestyle: P-value (ANOVA) ->6.88306902049e-10

EquipmentRental: P-value (ANOVA) ->0.0

EquipmentLastMonth: P-value (ANOVA) ->0.0

CallingCard: P-value (ANOVA) ->2.63562745643e-124

WirelessData: P-value (ANOVA) ->2.12831467911e-88

Multiline: P-value (ANOVA) ->0.0

VM: P-value (ANOVA) ->4.44294335801e-78

Pager: P-value (ANOVA) ->1.47627221103e-72

Internet: P-value (ANOVA) ->6.38901875529e-124

CallerID: P-value (ANOVA) ->1.97935888706e-06

CallWait: P-value (ANOVA) ->3.4838643101e-05

CallForward: P-value (ANOVA) ->6.2529979066e-05

ThreeWayCalling: P-value (ANOVA) ->9.65698924341e-07

EBilling: P-value (ANOVA) ->3.90338981268e-103

OwnsPC: P-value (ANOVA) ->1.43455024892e-58

OwnsMobileDevice: P-value (ANOVA) ->1.22730123045e-36

OwnsGameSystem: P-value (ANOVA) ->1.66095767326e-34

OwnsFax: P-value (ANOVA) ->1.36950961901e-64

NewsSubscriber: P-value (ANOVA) ->1.57978770248e-178

HouseholdIncome: P-value (ANOVA) ->5.39293865512e-60

Education: P-value (ANOVA) ->2.45951710915e-45

EquipmentOverTeunre: P-value (ANOVA) ->0.0